

食事と健康状態の関連を知るための一手法

李 丹陽* 高田 雅美* 城 和貴*

danyangl@ics.nara-wu.ac.jp

* 奈良女子大学大学院 人間文化研究科 情報科学専攻

概要

近年、科学進歩に伴い、記憶装置の大容量化が進んでいる。その結果として、蓄積された大量のデータの中から得られる情報は、多種多様かつ複雑である。そのため、従来の統計解析手法では扱うことが難しいデータや、様々な形式のデータベースから、有用な情報を取り出す必要がある。このための技術として、データマイニングが注目されている。本論文では、摂取した食品と健康状態のデータに対してデータマイニングを適用することで、摂取した食品と健康状態の関連について人間の先入観を介入させず発見し、健康状態の管理に役立つ指標を作るために、食事と健康状態の関連を調べる手順を提案する。また、データマイニングを用いて食事と健康状態の相関ルールを発見するための実験を行う。

A Case Study of Explore the Relationship Between Food and Health Condition

Danyang Li* Masami Takata* Kazuki Joe*

* Graduate School of Humanities and Sciences, Nara Women's University

Abstract

Recently, the size of memory storage is getting larger by advancement of science. As a result, information found in reservoir data has great variety and complexity. Therefore, it is necessary to get useful information from the data, which is hard to treat by the previous statistic analyses, or the database, which has variety of formats. To do so, the data mining is useful technique. This article tries to find the relationship between ingested food and condition of health without human prejudices and make new important indicator of controlling human health by employing data mining technique. Also the experiment is held to find the association-rules between food and condition of health.

1 はじめに

近年、科学進歩に伴い、記憶装置の大容量化が進んでいる。その結果として、蓄積された大量のデータの中から得られる情報は、多種多様かつ複雑である。そのため、従来の統計解析手法では扱うことが難しいデータや、様々な形式のデータベースから、有用な情報を取り出す必要がある。このための技術として、データマイニングが注目されている。データマイニングを用いることにより、集めたデータからなんらかの知見を発見することが期待される。

人間は、良好な健康状態を得るために、摂取する食べ物に関する情報を必要とする。食物と健康の関連を知るために、試験者のアンケート結果をもとに、データマイニングを行う。本論文では、食事と健康状態の関連を調べるための手順を提案する。また、データマイニングを用いて食事と健康状態の相関ルールを発見する実験を行う。

本論文の2章において、健康管理について述べる。3

章では、データマイニングを用いた食事と健康の管理システムに関する手法および実験環境について説明する。4章では、提案するシステムを実装し、実験結果について述べ、考察する。5章でまとめ。

2 健康管理について

生活習慣病の予防、健康維持・増進などを図るための栄養摂取量の研究や、健康的な食事摂取プランの開発の研究[1]がある。しかし、これらは食品と健康状態の直接的な関連に関する研究ではない。食事による摂取エネルギー、睡眠時間、飲酒量、喫煙量など生活習慣データと、血圧、体重、体脂肪率など健康状態データに関する相関ルール解析を行う健康データマイニングシステムの開発研究[2]もある。しかし、食品の摂取量、睡眠時間、飲酒量、喫煙量など生活習慣データが必要なため、被験者の負担が大きい。この負担を減らすためには、大まかな食品の摂取と健康状態のみの

データから、特徴を発見し、健康状態の把握や管理をすることが考えられる。このことにより、より日常的に簡単に食事と健康の関係を知ることができると考えられる。そこで、摂取した食品と健康状態のデータに対してデータマイニングを適用することによって、摂取した食品と健康状態の関連について人間の先入観を介入させず発見するためのシステムを構築すべきである。健康状態を知るためのパラメータとして、本研究グループでは、排泄物の形状から健康状態を推論するための一手法に関する研究も行っている。一方、本研究では、摂取した食品と健康状態に対するデータマイニングのシステム開発を行う。

3 データマイニングを用いた食事と健康の管理システム

データマイニングの代表的な手法として、アプリアリアルゴリズムがある。アプリアリアルゴリズムは、「長さ k の頻出でないパターンを含む長さ $k+1$ のパターンは頻出でない」という理論の元で、頻出パターンを抽出するアルゴリズムである [3]。本論文では、アプリアリアルゴリズムを用い、食事と健康状態の相関ルール [4] を発見する。さらに、相関ルールの評価基準としてカイ 2 乗検定 [5] を行う。今回、カイ 2 乗検定の有意水準 α を 5% とする。相関ルール $X \Rightarrow Y$ とする。 X , Y , $X \cup Y$ のサポートをそれぞれ S_X , S_Y , S_{XY} とし、トランザクションの総数を N とする。ここで X と Y が独立し、同じトランザクション内に含まれるのが単なる偶然であると仮定する。カイ 2 乗の検定量 T_{dep} は次のようになる。

$$T_{dep} = N \frac{(S_{XY} - S_X)^2}{S_X S_Y (1 - S_Y)(1 - S_X)}$$

検定量は自由度 1 のカイ 2 乗分布に従うことが知られている。カイ 2 乗の検定量 T_{dep} の値が 0 に近ければ X と Y はお互いに独立であり、大きければ相関が強いといえる。そこで、ある有意水準 α を定め、 $T_{dep} < x_1^2(\alpha)$ であれば X と Y が独立であると見なし、相関ルール $X \Rightarrow Y$ が発見されたのは単なる偶然であるから、価値がないとして捨てる。

4 実験

4.1 実験環境

今回実験用のデータは、健康によい食べ物、健康によくない食べ物の特徴を発見する題材として、ある基準を元に 198 品目のレシピを分析対象に選ぶ。これらは、肉類、魚介、野菜、豆腐、ご飯、めん、汁物、その他とおやつの 9 種類に分けられる。そして、20 代の女性 32 名をデータの対象者とする。

本論文の実験では、被験者が直接健康状態を入力したデータを利用することとする。実験に用いたデータセットは、1ヶ月に食べたレシピ履歴データスクリプトによって生成する。データの中では、日付、食べたレシピ、健康状態、便通の状態を記入されているが、今回はこの中から、日付、食べたレシピ、便通状態の 3 つのデータを切り取って使用した。表 1 に示すデータセットの一部の例を使って説明する。1 日に食べたレシピ、健康状態と便通状態を各行に表す。レシピ名の代わりに、あらかじめ付けられたレシピ番号で表す。例えば、表の最初の日に食べたレシピの中のレシピ番号 95 は、対応となるレシピ名は大根サラダじゃこドレッシングである。朝昼晩の順番でなく、レシピ番号の昇順で表す。ここでは、一日中に同じレシピを複数回食べることを認める。便通状態を把握するため、食事アンケートでは便通状態を適当に数値に変換して計算する。例えば、「便が出なかった」、「かたい便がでた」、「やわらかい便がでた」、「水状の便がでた」であれば、それぞれ 0, 1, 2, 3 と変換することをあらかじめ設定しておく。表の便通状態の欄では、対応となる番号を表示される。表 1 において、11/19 のデータがチェック漏れや記入の忘れと見なす。11/21 のデータのレシピ番号の欄では、0 を記入されることは何も食べていないことを表す。

便秘とは一般的に、排便が順調に行われない状態のことを言う。しかし、排便の回数には個人差が大きく、便秘がある期間の排便回数で定義することは非常に難しい。今回の実験では、1 日に排便の回数が 0 であれば、便秘と見なす。

4.2 実験結果

32 人の 1ヶ月のデータを用い、便秘の前日の 1 日分のレシピのみからなるデータ集合に着目して単体および組み合わせでデータマイニングを行う。また、便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと考え、便秘の前の複数日間分のレシピからなるデータ集合に着目して組み合わせでデータマイニングを行う。

まず、便秘の前日の 1 日分のレシピのみからなるデータ集合に着目する。この場合では、相関ルールの前提部 X をレシピ A とし、相関ルールの結論部 Y を次の日に便秘になることとする。最小確信度を 50% とし、カイ 2 乗分布の有意水準を 5% とする。また、1 日のデータを一つのトランザクションとする。データセットにおいて、各人に対してそれぞれの便通状態が 0 であれば、対応となる前日のレシピを取り出す。各人のデータの中では、連続していないデータがある（例えば、記入漏れや何も食べていない日など）。このため、便秘の前日が存在しない場合は無視する。抽出された相関ルールのカイ 2 乗検定量 T_{dep} を求める。求められたカイ 2 乗検定量の値は（カイ 2 乗分布表より有意水準 5% の時、検定量 T_{dep} の値が 3.841 となる）3.841

日付	レシピ番号								健康	便通
11/15	95	95	138	145	186	701	706		1	2
11/16	125	138	138	145	186	701	707		1	2
11/17	118	131	138	138	145	186	701		1	2
11/18	103	138	145	146	701				1	2
11/20	50	83	120	128	138	145	701		1	2
11/21	0								1	1
11/22	137	137	145	186	701	702	706		1	2
11/23	35	136	138	145	186	701			1	2
11/24	138	145	145	177	183	701	705		1	0

表 1: データセットの一部

レシピ	検定量	確信度	サポート
酢豚	6.849704	0.833333	0.006361
あさりの酒蒸し	4.033251	1	0.002545
関西風雑煮	4.033251	1	0.002545
大豆とちりめんじゃこのいり煮	7.856304	0.727273	0.010178

表 2: レシピ単体の実験結果

レシピの組み合わせ	検定量
ハムとチーズのサンドイッチ, サクッとチョコレート	3.917093

表 3: 前日のみのレシピの組み合わせの実験結果

より大きい相関ルールの前提部と結論部の相関が強い
ため、価値のある相関ルールとして出力される。

実験結果は表 2 に示す。次に、便秘の前日の 1 日分の
レシピのみからなるデータ集合に着目する。今回の
データセットは 32 人分の 1ヶ月のデータを用い、レ
シピが全部で 198 個あるため、同じレシピの出現確率
が非常に少ないので、最小確信度設定せずに、最小サ
ポートを 20 とする。1 日のデータを 1 つのトランザク
ションとする。この場合、相関ルールの前提部 X を便秘
の前日に食べたレシピの組み合わせとし、次の日に
便秘となることを結論部 Y とする。実験結果は表 3 に
示す。同様に、2 日前, 3 日前, 4 日前の実験結果がそ
れぞれ表 4, 5, 6 に示す。

さらに、便秘の前の複数日間分のレシピからなるデー
タ集合に着目して組み合わせでデータマイニングを行
う実験を述べる。ここでは、便秘の前の複数日間分を
 N とする。 $2 \leq N \leq 4$ について実験を行う。まず、便秘
の前 2 日間分のデータを一つのトランザクションと
する場合について述べる。この場合、相関ルールの前
提部 X を便秘の前 2 日間分のレシピの組み合わせとし、
2 日後に便秘となることを結論部 Y とする。実験
結果は表 7 に示す。同様に、前 3 日間分と前 4 日間分
のレシピの組み合わせの実験結果が得られる。表 8 と
表 9 に示す。

4.3 考察

アプリアリアルゴリズムによって、食事と健康状態
に関するデータに対してデータマイニングを行う実験
では、便秘の前日の 1 日分のレシピのみからなるデー

レシピの組み合わせ	検定量
だし巻き卵, 白粥	8.15726
白粥, ハムとチーズのサンドイッチ	5.02715
ハムとチーズのサンドイッチ, サクッとチョコレート	10.866193

表 4: 2 日前のみのレシピの組み合わせの実験結果

レシピの組み合わせ	検定量
鶏のから揚げ, 白粥	5.392515
だし巻き卵, 白粥	11.206815
白粥, サクッとチョコレート	3.889798
ハムとチーズのサンドイッチ, サクッとチョコレート	20.681059

表 5: 3 日前のみのレシピの組み合わせの実験結果

タ集合に着目して単体および組み合わせでデータマイ
ニングを行うことと、便秘の前の複数日間分のレシピ
からなるデータ集合に着目して組み合わせに着目する。

まず、レシピの単体の場合について考察する。酢豚、
大豆とちりめんじゃこのいり煮、あさりの酒蒸しと関
西風雑煮を食べると、次の日に便秘になる可能性が高
いと予測することができる。

次に、レシピの組み合わせの場合について述べる。
まず、便秘の前日の 1 日分のレシピのみからなるデー
タ集合に着目して組み合わせでデータマイニングを行
う実験についての考察を述べる。便秘の前日のみに着
目すると、ハムとチーズのサンドイッチとサクッとチ
ョコレートの組み合わせを食べると、次の日に便秘に
なる可能性が高いといえる。便秘の 2 日前の場合では、
出し巻き卵と白粥、白粥とハムとチーズのサンドイ
ッチ、そして、ハムとチーズのサンドイッチとサクッ
とチョコレートの 3 つの組み合わせのどれかを食べ
ると、2 日後に便秘になる可能性が高いと予想するこ
とができる。また、便秘の前の 2 日間分のレシピから
なるデータ集合に着目し、組み合わせでデータマイ
ニングを行う実験結果と合わせて考察すると、便秘の
前日と 2 日前共にハムとチーズのサンドイッチとサ
クッとチョコレートの組み合わせという結果になり、
便秘の前の 2 日間の結果に影響していると考えられ
る。次に、便秘の 3 日前の実験結果と便秘の前の 3
日間分の実験結果から見ると、白粥、ハムとチー
ズのサンドイッチとサクッとチョコレートの 3 つの
組み合わせのうちのどれかを食べると、3 日後に
便秘になる可能性が高いと考えられる。最後に、
便秘の 4 日前の実験結果が白粥、ハムとチー
ズのサンドイッチとサクッとチョコレート

レシピの組み合わせ	検定量
だし巻き卵、白粥	11.222124
ハムとチーズのサンドイッチ、サクッとチョコレート	19.64843
白粥、ハムとチーズのサンドイッチ、サクッとチョコレート	4.101117

表 6: 4 日前のみのレシピの組み合わせの実験結果

レシピの組み合わせ	検定量
ハムとチーズのサンドイッチ、サクッとチョコレート	4.876865

表 7: N=2 の場合のレシピの組み合わせの実験結果

の組み合わせの他に、出し巻き卵や鶏のから揚げの出現回数が高いにも関わらず、便秘の前の 4 日間分の実験結果では、白粥、ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせが食べると、4 日後に便秘になる可能性が高いという結果から、便秘の前の 4 日間分をまとめて見ることによって、4 日後に便秘になることに関する予測が不十分であることがわかる。以上より、ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせを食べると、便秘になる可能性が高いと予想することができると考えられる。また、便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと言える。

レシピの単体の場合とレシピの組み合わせの場合では、全然違う結果が得られた。これは、レシピの単体の場合で得られた実験結果のレシピの組み合わせは、全体の中での出現回数、即ちサポートが少なく、設定された最小サポート 20 以下であるため、アプリオリアルゴリズムによって枝刈りされたためだと考えられる。

今回の実験で使用されたレシピの項目には偏りがある。例えば、果物、ヨーグルトなどいわゆる便秘の改善に良い食べ物と、サプリメントや薬など便秘に影響するレシピがないことである。そのため、限られたレシピで実験を行ったため、調査として十分とは言えないところがある。また、一日に数回排便がした場合の考慮をしていないことと、便の固さの判断の個人差も考慮せずに実験を行ったため、得られた結果も多少のノイズを含んでいると考えられる。また、食べ物の順序を考慮したマイニングは行っていないこと、データ数が不足していること、レシピの数が多いなども結果に多く影響する原因と考えられる。また、便秘に影響する月経も一つの要素として挙げられる。従って、実験で得られた結果はあくまでも予測としか言えないと考える。

5 まとめ

本論文では、データマイニングを用いて食事と健康状態の関連の調べに関する手順を提案し、実験を行った。今回の実験の結果、レシピの単体の場合では、酢豚、アサリの酒蒸し、関西風雑煮と大豆とちりめんじゃこのめんじゃこのいり煮を食べると、次に日に便秘に

レシピの組み合わせ	検定量
白粥、ハムとチーズのサンドイッチ	4.3198314
白粥、おからうどん	3.900086
白粥、サクッとチョコレート	4.993071
ハムとチーズのサンドイッチ、サクッとチョコレート	6.900366

表 8: N=3 の場合のレシピの組み合わせの実験結果

レシピの組み合わせ	検定量
白粥、ハムとチーズのサンドイッチ	5.176787
ハムとチーズのサンドイッチ、サクッとチョコレート	4.040732

表 9: N=4 の場合のレシピの組み合わせの実験結果

なる可能性が高いと予想することができる。また、レシピの組み合わせの場合では、ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせを食べると、次の日、二日後、三日後と四日後に便秘になる可能性が大きいと予想することができた。

今後、より確信度高い予想を得られるため、データマイニングの対象とするデータベースの形式をレシピより、素材まで分類する必要があると考えられる。また、長期的にデータを蓄積する場合、アプリオリアルゴリズムの記憶量が大きく必要となるため、新たな手法を用いてマイニングする必要があると考えられる。今回は、ホームページからデータ入力を行い、コンピュータ上でデータマイニングを行っている。今後は、データ入力とデータマイニングを行うアプリを開発したいと考えている。

参考文献

- [1] 食事摂取プランの研究:
http://www.v350f200.com/kanri/kankei_2g.html
- [2] 竹内裕之, 児玉直樹, 橋口猛志, 林同文”個人健康管理を目的とした健康データマイニングシステム”, DEWS2006 論文集, 1B-ill, 2006
- [3] R.Agrawal, A.Arning, T.Bollinger, M.Mehta, J.Shafer, and R.Srikant, The Quest data mining system. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1996
- [4] Ian H.Witten, Eibe Frank, "Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations" Morgan Kaufmann Publishers, 1999
- [5] 福田剛志, 森本康彦, 徳山豪”データマイニング”共立出版, 2001